Schweihs_8_r_3

Maggie Schweihs

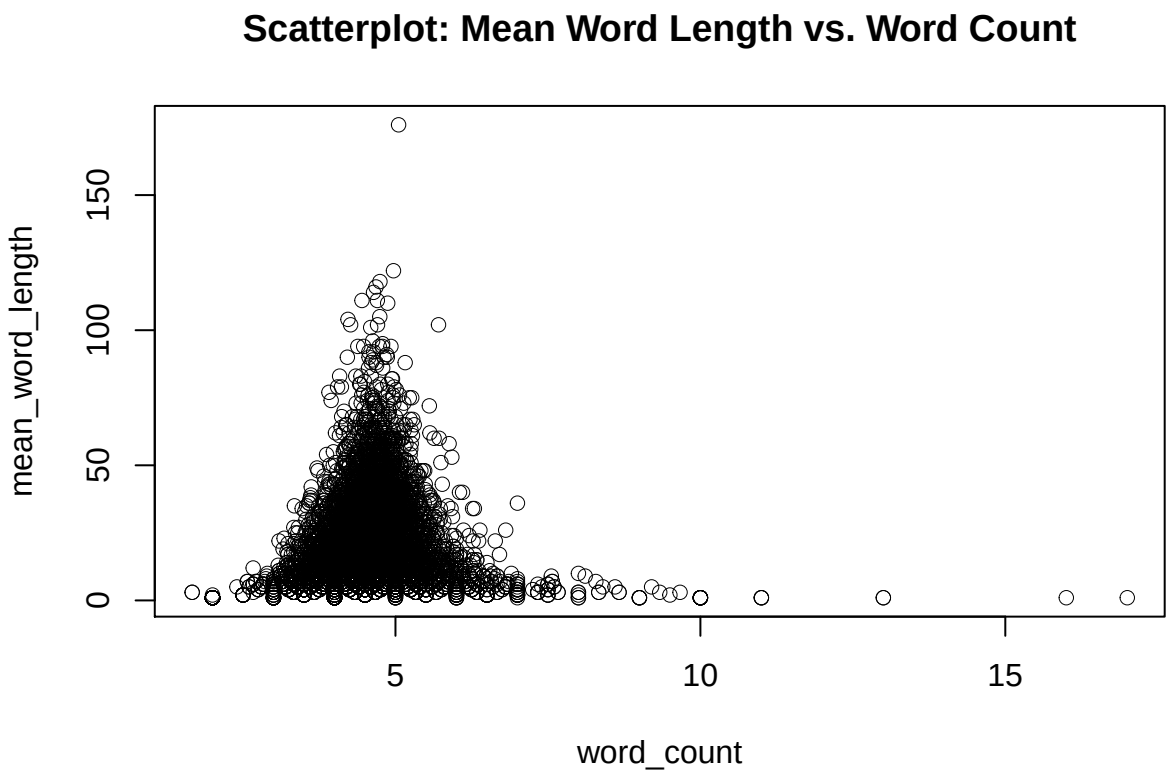
10/29/2016

3. Sentence Length v. Word Length

Investigate whether sentences with more words tend to contain longer words.

3.a: Load the data set into R and make a scatterplot of mean length of words versus number of words per sentence.

```
pride <- read.csv("~/Documents/Code/DS710/ds710fall2016assignment7/pride.csv", header = TRUE)
attach(pride)
plot(mean_word_length ~ word_count, type = "p", lwd = .4, main = "Scatterplot: Mean Word Length vs. Wor
```



Does Linear Regression make sense here?

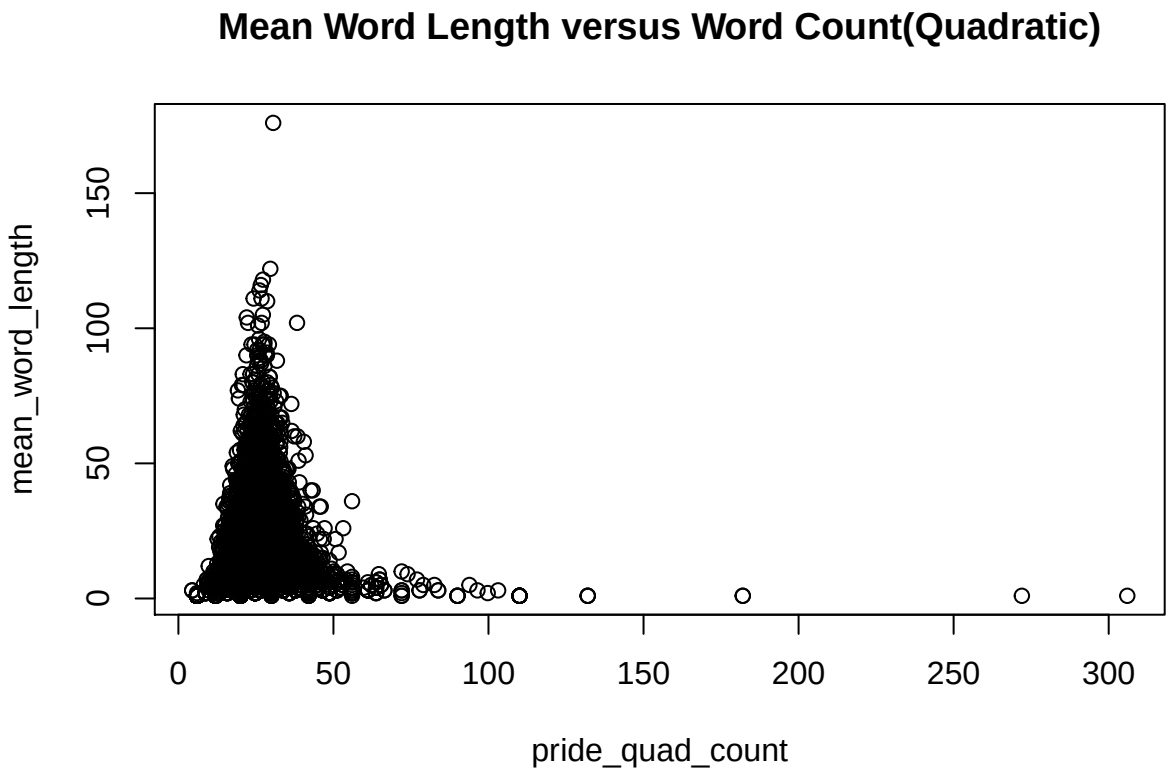
Looking at the scatterplot of mean length of words versus number of words per sentence, the data appears to have a polynomial relationship, due to one or more noticable humps on the graph. Linear regressions **would not** be appropriate here.

3.b: Trial and Error

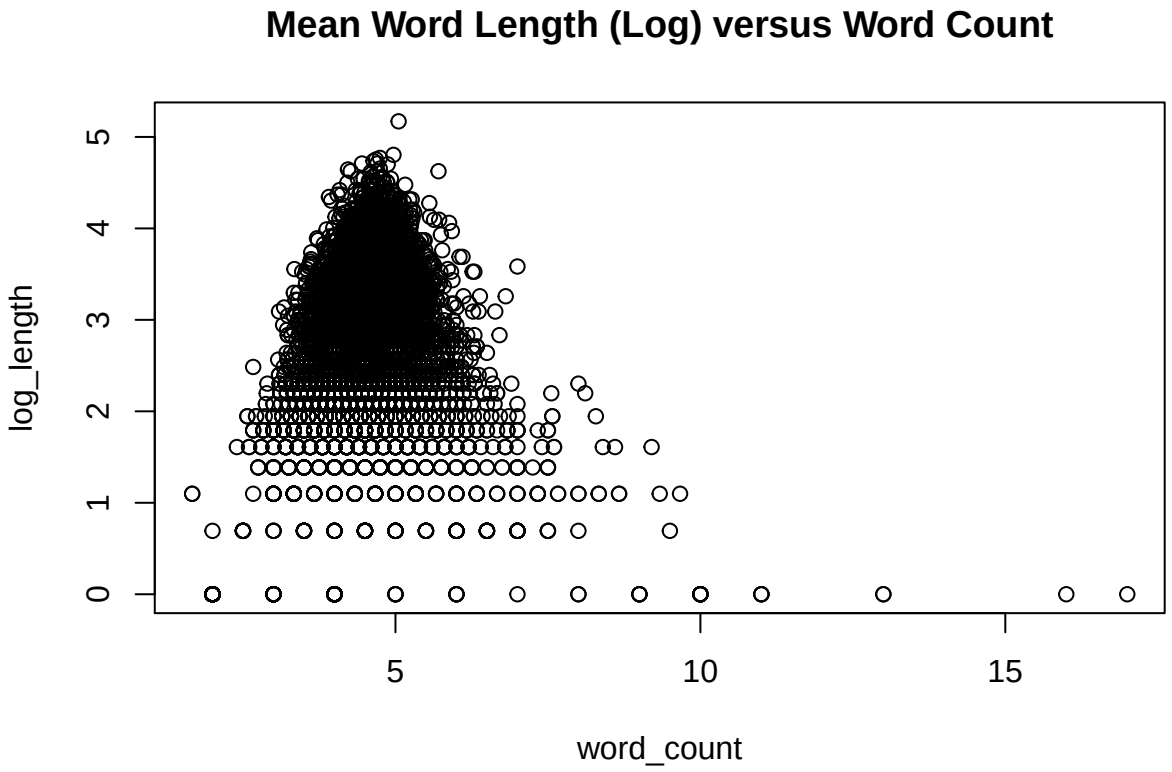
For lack of a better method, I tried various combinations of variables and transformations of variables: logarithmic transformation, quadratic functions, cubic functions, and combinations therein. Below is a subset of that trial and error. Along the way I did **Not** wind up with anything appearing to be linear.

```r
# For Lack of knowledge of a statistical tool to measure and compare
# relationships between variables, I resorted to good ole trial and error

#Initializing some variables to store the functions and transformations
pride_cubic_length <- (mean_word_length + (mean_word_length)^3 +(mean_word_length)^2)
pride_cubic_count <- (word_count + (word_count)^3 +(word_count)^2)
pride_quad_count <- (word_count +(word_count)^2)
pride_quart_length <- ((mean_word_length)^4 + (mean_word_length)^3 +(mean_word_length)^2 + mean_word_le
pride_quad_length <- (mean_word_length +(mean_word_length)^2)
log_length <- log(mean_word_length)
log_count <- log(word_count)
#plot some Combinations
plot(mean_word_length~pride_quad_count, main = "Mean Word Length versus Word Count(Quadratic)" )
```




```r
plot(log_length~word_count, main = "Mean Word Length (Log) versus Word Count" )
```



Neither one of the above transformations yielded a plot appropriate for linear regression

3.c: c. Test whether sentences with more words tend to contain longer words. State your conclusion in context.I'm going to test the relationship between `mean_word_length` and the quadratic word count function.

H_0: Longer sentences do not typically have longer words in Pride and Prejudice. mu = 0

H_1: Longer sentences **do** have longer words in Pride and Prejudice. mu != 0

Test using the significance test for linear regression.

```
pride.lm <- lm(mean_word_length~pride_quad_count)
pride.lm
```

```
##
## Call:
## lm(formula = mean_word_length ~ pride_quad_count)
##
## Coefficients:
##      (Intercept)  pride_quad_count
##        19.989871          0.006961
```

```
summary(pride.lm)
```

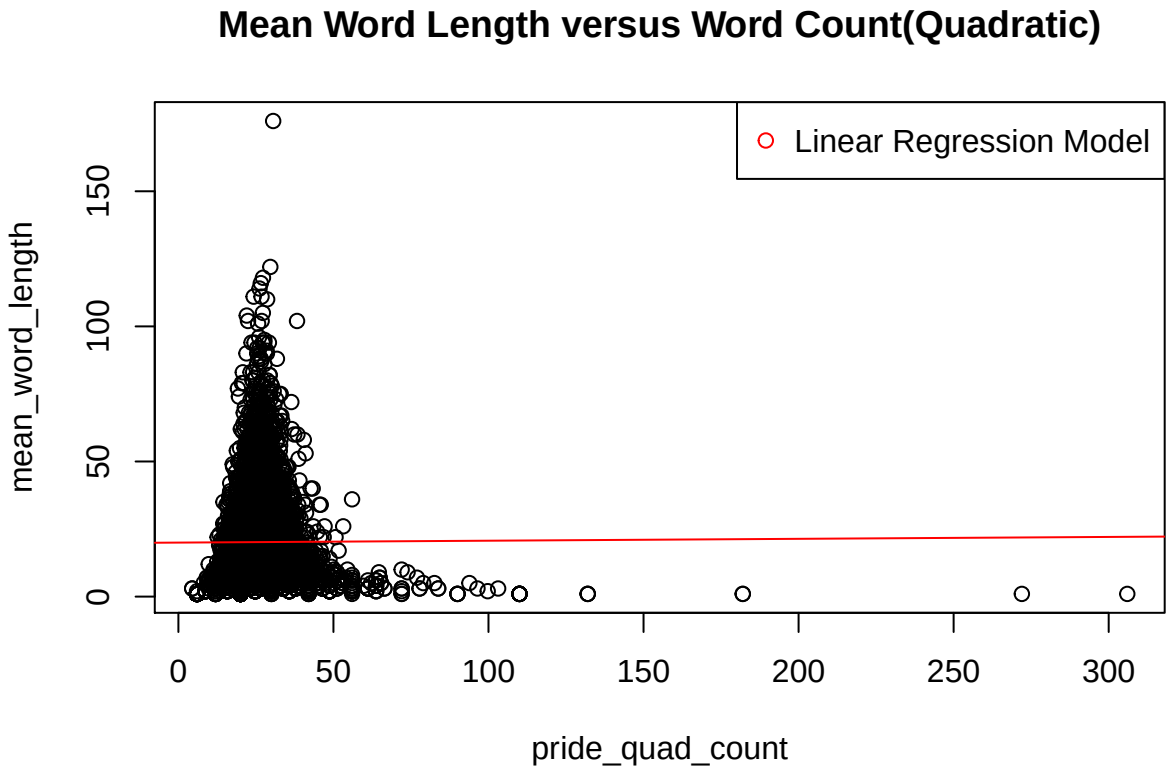
```
##
## Call:
## lm(formula = mean_word_length ~ pride_quad_count)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.120 -12.114  -4.204   7.773 155.797
```

```
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      19.989871   0.542086  36.876   <2e-16 ***
## pride_quad_count  0.006961   0.019160   0.363    0.716
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.29 on 6199 degrees of freedom
## Multiple R-squared:  2.129e-05,  Adjusted R-squared:  -0.00014
## F-statistic: 0.132 on 1 and 6199 DF,  p-value: 0.7164
```

The p-value for the test is p=0.7164, so we **cannot** reject the null hypothesis. There is not significant evidence here to suggest a relationship between mean word length and the number of words in a sentence.

3.d: Add a line to my scatterplot representing the regression model.

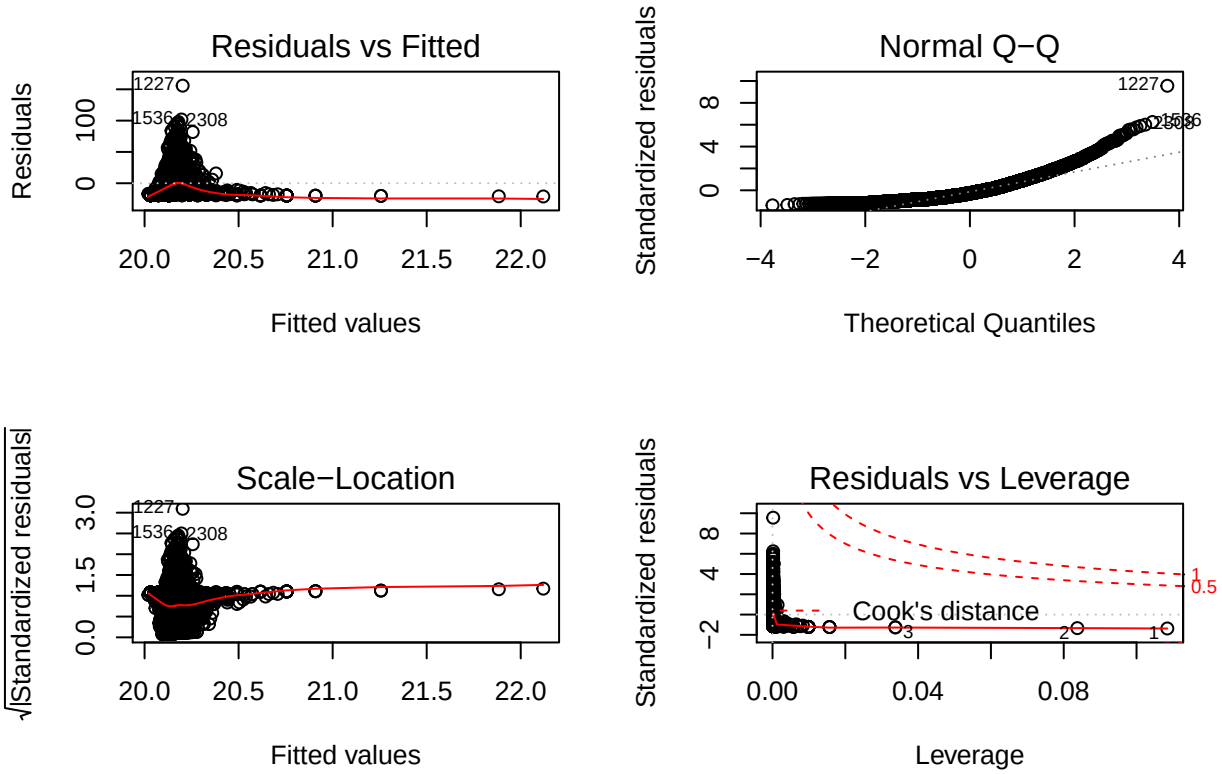
```
plot(mean_word_length~pride_quad_count, main = "Mean Word Length versus Word Count(Quadratic)" )
abline(lm(mean_word_length~pride_quad_count), col = "red", cex =1)
legend("topright", c("Linear Regression Model"), col ="red", pch = 21)
```



The slope of the linear regression line is roughly 0.007 and is shown to be non-significant. We can see this by observe the graph, as well. The slope of the line is nearly zero, which reaffirms the result of the significance test: we did not show a linear relationship between the variables with this model. Recommendation: ecplore non-linear models.

3.e: Examine the residual diagnostic plots, and explain what they tell us in this case.

```
par( mfrow = c( 2, 2 ) )
plot(pride.lm)
```



By looking at the Residuals versus Fitted plot, we reaffirm the intuition that some affects of the dependent variable are not being taken into account by our model.The Normal Q-Q plot suggests that our data is right-skewed and possibly bi-modal.